Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

1. The target variable *cnt* that represents total bike sharing rentals increases year over year.
2. It is highest in the months of April to October
3. It is significantly higher on non-holidays
4. It is high on Summer and Fall
5. It is high on non-rainy days.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

When we convert a categorical variable into dummy variables, each category is represented as a binary variable (0 or 1). If all dummy variables for a category are included, they are perfectly correlated with each other. This causes multicollinearity, leading to issues in regression analysis, such as unstable coefficients and unreliable p-values. To avoid this, one of the dummy variables is dropped (typically the first or reference category). This makes the remaining variables independent and avoids redundancy. When you set *drop_first=True*, it automatically drops one dummy variable for each categorical column, using it as the reference category.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

The target variable cnt has the highest correlation with temperature specific predictor variables `temp` and `atemp`. The correlation value for both with cnt is 0.63.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

1. Linearity – Plotted a scatterplot graph of y_test (actual values) and y_test_pred (predicted values) and also y_train and y_train_pred and confirmed that it is roughly linear.
2. Normal Distribution of error terms - Plotted the histogram of the error terms and verified that it is normally distributed
3. Constant Variance in Error Terms (Homoscedasticity) - Plotted a graph of residuals vs predicted values and observed that the residuals scatter **randomly and evenly** around the horizontal axis (y = 0). There is **no clear pattern** or systematic change in the spread of residuals. This indicates that the variance of residuals is constant across all levels of predicted values.
4. No Multicollinearity – Calculated VIF (Variance Inflation Factor) of the predictor variables

and ensured that the variables are not highly correlated with each other. Whenever there were high values, made corrections accordingly by removing the high correlated variables.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)
The top three features that contribute significantly towards demand of the shared bikes are –

1. temp
2. yr
3. winter

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear regression is a statistical method used to model the relationship between one or more independent variables (X) and a dependent variable (y). The goal is to find a linear equation that best predicts y bar based on X bar.

The linear regression hypothesis function is:
y= β0 + β1x1 + β2x2 +…+ βnxn + ϵ

Where:

y : Dependent variable (target).
x1, x2, …, xn: Independent variables (features).
β0: Intercept (constant term).
β1, β2, …, βn: Coefficients (slopes) for the features.
ϵ: Error term (captures unexplained variability).

The objective of linear regression is to minimize the differences between the predicted values (y^) and the actual values (y) by estimating the coefficients (βx).

Linear regression uses the Mean Squared Error (MSE) as its loss function:

MSE = 1/m * Sigma(i=1 to m) (yi - y^i)^2

Where:
m: Number of data points.
yi: Actual value.
y^i : Predicted value

Minimizing the MSE ensures that the predicted values are as close as possible to the actual values.

Algorithm Steps:

1. Initialize Parameters
   - Start with random initial values for β0, β1, …, βn
2. Fit the Line (Ordinary Least Squares - OLS)
   - Find the values of β that minimize the MSE using the formula:

$B = (X^T X)^{-1} X^T y$

Where:

- X: Matrix of feature values, including a column of 1s for the intercept.
- y: Vector of actual target values.

OLS provides the best-fit line in the sense of minimizing the squared differences between observed and predicted values.

3. Prediction

Once the coefficients (β) are determined, predictions can be made using:

$y\hat{} = \beta 0 + \beta 1x1 + \ldots + \beta n$

4. Model Evaluation

Evaluate the model's performance using metrics such as:

- R^2: Proportion of variance explained by the model.
- Adjusted R^2: Adjusted for the number of predictors.
- RMSE (Root Mean Squared Error): Square root of MSE.

Assumptions of Linear Regression

1. Linearity: The relationship between XXX and yyy is linear.
2. Homoscedasticity: Constant variance of errors.
3. Normality of Errors: Errors are normally distributed.
4. No Multicollinearity: Predictors are not highly correlated with each other

---

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's Quartet is a set of four distinct datasets created by statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analyzing it. Despite having nearly identical basic statistical properties, these datasets have very different distributions and relationships when plotted. This highlights the risk of relying solely on summary statistics for data analysis.

Characteristics of Anscombe's Quartet

Each dataset in the quartet consists of 11 (x,y) pairs. The datasets share the following nearly identical properties:

1. Mean of x and y:
   - Mean of x=9.0
   - Mean of y=7.5
2. Variance of x:
   - Variance of x=11.0
3. Regression Line:
   - The linear regression line for each dataset is y=3+0.5x

4. Correlation Coefficient:
    - r=0.816
5. Sum of Squares of Residuals:
    - Identical across all datasets.

Despite these similarities, the datasets are visually and structurally very different.

The Four Datasets

1. Dataset 1: A Typical Linear Relationship
    - The data fits the regression line well.
    - Scatterplot shows a linear relationship between x and y.
2. Dataset 2: A Non-linear Relationship
    - The data points form a curved pattern.
    - A linear regression line is inappropriate for this data.
3. Dataset 3: Outlier-Driven Relationship
    - Most data points align well with the regression line, but one extreme outlier heavily influences the slope.
    - The correlation is misleading due to the outlier.
4. Dataset 4: Vertical Cluster with One Outlier
    - All x-values except one are identical.
    - The single outlier determines the slope of the regression line.
    - The linear relationship is an artifact of the single point.

Key Takeaways

1. Importance of Visualization:
    - Summary statistics like mean, variance, correlation, and regression lines can hide critical differences in the data.
    - Always visualize your data to identify patterns, outliers, and non-linear relationships.
2. Outliers Matter:
    - Outliers can heavily influence statistical measures like the correlation coefficient and regression slope.
3. Context is Key:
    - Statistical properties alone cannot tell the full story; interpreting results requires understanding the data's context.

Lessons for Data Analysis

1. Always visualize data, especially when working with small datasets.
2. Use visualization to check for assumptions of regression and identify anomalies.
3. Relying only on summary statistics can lead to incorrect conclusions.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is one of the most widely used correlation coefficients.

Formula

Pearson's rrr is calculated as:

$r=Cov(X,Y)/\sigma X\sigma Y$

Where:

- Cov(X,Y): Covariance between variables X and Y.
- σX : Standard deviation of X.
- σY : Standard deviation of Y.

Alternatively, in a computational form:

$r=\sum(xi-\bar{x})(yi-\bar{y})/sqrt(\sum(xi-\bar{x})^2\cdot\sum(yi-\bar{y})^2)$

Where:

- xi, yi : Data points for X and Y.
- $\bar{x},\bar{y}$: Means of X and Y.

Interpretation

The value of r ranges between −1 and +1:

1. r = +1: Perfect positive linear correlation (as X increases, Y increases).
2. r = −1: Perfect negative linear correlation (as X increases, Y decreases).
3. r = 0: No linear correlation (no linear relationship between X and Y).

Assumptions of Pearson's r

1. Linearity:
    - Assumes a linear relationship between X and Y.
    - It does not work well for non-linear relationships.
2. Continuous Variables:
    - Both X and Y should be continuous (interval or ratio scale).
3. Normality:
    - Ideally, X and Y should be approximately normally distributed.
4. Homogeneity of Variance:
    - The spread of Y values should be roughly the same across all X values.
5. No Outliers:
    - Outliers can distort the value of r.

Uses

- Quantifying the relationship between two variables (e.g., height and weight, temperature and energy consumption).
- Identifying potential predictors for regression analysis.
- Testing hypotheses about relationships between variables.

Limitations

1. Linear Relationships Only:
    - Pearson's r only measures linear relationships. It may give r=0 even if there is a strong non-linear relationship.
2. Sensitivity to Outliers:
    - Outliers can significantly skew the correlation coefficient.
3. Correlation is Not Causation:
    - A strong correlation does not imply that one variable causes changes in the other.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling is a data preprocessing technique used to adjust the range of features (variables) in a dataset so that they can be compared on the same scale. It transforms the data without changing its structure, ensuring that all features contribute equally to the model.

Why is Scaling Performed?

1. To Improve Model Performance:
    - Many machine learning algorithms are sensitive to the scale of input features. Features with larger ranges can dominate features with smaller ranges, leading to biased results.
2. To Optimize Distance-Based Algorithms:
    - Algorithms like k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), and clustering methods rely on distance calculations. Scaling ensures that one feature does not disproportionately influence the results.
3. To Speed Up Convergence:
    - Gradient-based algorithms (e.g., logistic regression, neural networks) converge faster when features are scaled because the gradient steps are uniform across dimensions.
4. To Address Feature Variance:
    - Features with larger variance can overpower those with smaller variance. Scaling ensures features are comparable.

Types of Scaling

1. Standardization (Z-Score Scaling)

- Definition:
    - Standardization rescales the data to have a mean of 0 and a standard deviation of 1.
- When to Use:
    - When features follow a Gaussian (normal) distribution.
    - When algorithms assume standardized input (e.g., PCA, logistic regression).
- Example: A feature with values [50, 60, 70] would be transformed into [-1.22, 0, 1.22] if its mean is 60 and standard deviation is 8.16.

2. Normalization (Min-Max Scaling)

- Definition:
    - Normalization rescales the data to fit within a fixed range, typically [0, 1].
- When to Use:
    - When you know the data doesn't follow a Gaussian distribution.
    - When features need to be bounded (e.g., for neural networks).
- Example: A feature with values [10, 20, 30] would be transformed into [0, 0.5, 1] if its minimum is 10 and maximum is 30.

Differences Between Standardization and Normalization

Feature	Standardization (Z-Score)	Normalization (Min-Max)
Range	No fixed range (can be negative or positive).	Scales data to a fixed range (usually [0, 1]).
Focus	Centers data by subtracting the mean and scaling to unit variance.	Scales data relative to min and max values.
When to Use	Data follows Gaussian distribution; algorithms like PCA, SVM.	Data does not follow Gaussian distribution; bounded outputs required.
Effect on Outliers	Sensitive to outliers (but less than normalization).	Highly sensitive to outliers.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

The formula for VIF is $1/(1-R_i^2)$. Infinite VIF means denominator is 0. The denominator can become zero when R_i is +1 or - 1. R_i = +1 or -1 indicates perfect correlation between ith predictor and other predictors.

What Causes Perfect Multicollinearity?

- Duplicate Columns:
  - If a variable is repeated in the dataset (e.g., X_1 and X_2 are identical).
- Linear Dependencies:
  - If one variable can be exactly predicted by a combination of others e.g., X3= X1+ X2
- Dummy Variable Trap:
  - Including all dummy variables for a categorical feature without dropping one category.
- Constant Features:
  - A variable with the same value across all observations adds no information but contributes to perfect correlation with the intercept.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. It helps assess whether the data follows the expected distribution.

How a Q-Q Plot Works

1. Axes:
    - X-axis: Theoretical quantiles from the reference distribution (e.g., a normal distribution).
    - Y-axis: Actual quantiles from the dataset.
2. Construction:
    - The data is ordered and plotted against the corresponding quantiles of the theoretical distribution.
    - A reference line (a 45-degree line) is included to represent perfect agreement between the data and the theoretical distribution.
3. Interpretation:
    - Points near the line: Data is well-approximated by the theoretical distribution.
    - Deviations from the line: Indicates departures from the theoretical distribution, such as skewness, kurtosis, or other anomalies.

Use and Importance of a Q-Q Plot in Linear Regression

In linear regression, the Q-Q plot is primarily used to assess the assumption of normality of residuals, which is crucial for valid statistical inference.

Why Check for Normality of Residuals?

1. Hypothesis Testing:
    - The normality of residuals is required for valid p-values and confidence intervals in linear regression.
2. Model Diagnostics:
    - Normal residuals ensure the model adequately fits the data without systematic patterns or biases.

How to Use a Q-Q Plot for Linear Regression?

1. Residual Analysis:
    - After fitting the regression model, extract the residuals (differences between actual and predicted values).
    - Create a Q-Q plot of the residuals against a normal distribution.
2. Interpretation in Linear Regression:
    - Points lie close to the line: Residuals are approximately normal.
    - S-shaped pattern: Indicates skewness in residuals.
    - Extreme deviations at ends: Suggests heavy tails or outliers.